DOCUMENT RESUME

ED 389 755 TM 024 388

AUTHOR Berger, Martijn P. F.; Veerkamp, Wim J. J.
TITLE A Review of Selection Methods for Optimal Test Design. Research Report 94-4.
INSTITUTION Twente Univ., Enschede (Netherlands). Faculty of Educational Science and Technology.
PUB DATE Nov 94
NOTE 39p.
AVAILABLE FROM Bibliotheek, Faculty of Educational Science and Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Adaptive Testing; Foreign Countries; *Research Design; *Selection; *Test Construction; Test Format; Test Items; Test Theory
IDENTIFIERS *Testlets

ABSTRACT

The designing of tests has been a source of concern for test developers over the past decade. Various kinds of test forms have been applied. Among these are the fixed-form test, the adaptive test, and the testlet. Each of these forms has its own design. In this paper, the construction of test forms is placed within the general framework of optimal design theory. A review of various objective functions and methods for the designing of different test forms is given. The advantages of using these methods are discussed, and an illustration of an optimal test design is provided. (Contains 3 figures, 1 table, and 36 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 389 755

T

A Review of Selection Methods for Optimal Test Design

Research Report 94-4

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. NELISSEN

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Martijn P.F. Berger

Wim J.J. Veerkamp

BEST COPY AVAILABLE

faculty of
EDUCATIONAL SCIENCE AND TECHNOLOGY

University of Twente

Department of Educational Measurement and Data Analysis

TM024388

2

A Review of Selection Methods for Optimal Test Design

Martijn P.F. Berger
Wim J.J. Veerkamp

To appear in G. Engelhard, Jr. & M. Wilson (Eds.), Objective Measurement: Theory into Practice (Vol. 3). Norwood, NJ: Ablex.

A review of selection methods for optimal test design, Martijn P.F. Berger and Wim J.J. Veerkamp - Enschede: University of Twente, Faculty of Educational Science and Technology, November 1994. - 33 pages.

Abstract

The designing of tests has been a source of concern for test developers over the past decade. Various kinds of test forms have been applied. Among these are the fixed-form test, the adaptive test and the testlet. Each of these forms has its own design. In this chapter the construction of test forms is placed within the general framework of optimal design theory. A review of various objective functions and methods for the designing of different test forms is given. The advantages of using these methods are discussed, and an illustration of an optimal test design will be given.

Key words: optimal test design, adaptive tests, testlets, sequential procedure, efficiency, consistency.

A Review of Selection Methods for Optimal Test Design

Since the First World War, the construction of tests in education and psychology has gone through a number of different stages, and tests have been administered in various different forms. Although at first the construction of tests was done by hand, the recognition that the construction of tests could be improved by taking into account the psychometric characteristics of the items has lead to alternative, and more structured methods of test construction. Perhaps one of the most promising directions in the construction of tests is the use of the idea of so-called item banks. An item bank is a very large set of items. These items are grouped into certain content areas and it is assumed that the psychometric characteristics of these items have been estimated. When such an item bank is available, the construction of a test is done by selecting items from the bank according to certain specifications. A lot of research has been done on optimal item selection methods. Many of these methods are based on mathematical programming procedures. See Adema (1990), Boekkooi-Timminga (1989), and Theunissen (1986) for a review of these methods. Although the mathematical programming methods were mainly proposed for the construction of fixed-form tests, other forms like two-stage and parallel tests (Adema, 1990) can also be handled by these methods. Recently two computer programs based on mathematical programming algorithms have been developed; namely the CONTEST program (Boekkooi-Timminga and Sun, 1991), and the OTD program (Verschoor, 1991).

The fact that fixed-form tests do not have equal reliability or equal

validity over the whole range of abilities in the population, has motivated Lord (1971, 1980) and Weiss (1976,1978), among others, to propose adaptive test forms. The central idea was that if each examinee in a sample is given an individually designed test, this would lead to more efficient estimation of the abilities of these examinees. The availability of fast computers and item response theory (IRT) models has made the development of computerized versions of adaptive testing (CAT) possible. See Wainer (1990) for a review of various aspects of CAT.

With the development of item banks and computerized adaptive tests, the special skills of the test developer were replaced by statistical characteristics. This development was criticized by Wainer and Kiely (1987). They argued that the test developer's skills are still needed in the construction process. Because several practical problems with the existing CAT procedures were not solved satisfactory, Wainer and Kiely (1987) and Wainer and Lewis (1990) proposed the application of so-called testlets. Testlets are actually small bundles of items, where examinees follow a fixed number of paths. A test may consist of a number of different testlets, and an examinee does not have to take every testlet in the test nor does an examinee have to take all items within the testlet. The many advantages and disadvantages of fixed-tests and adaptive tests are combined in a testlet design.

The above described construction of different test forms can be regarded as an optimal design problem. Optimal design methods have been applied in various fields of research. Although most of the developments have been reported and applied in bioassay research, optimal design methods can also be applied in educational measurement. Berger (1991) and Berger and van der Linden (1992), for example, have recently described the application of optimal design methods

for the designing of optimal samples for item parameter estimation in IRT models.

The objective of this paper is to give a review of the different optimal design methods and criteria for item selection for the construction of different test forms. This review places these methods within the general framework of optimal design theory. Silvey (1980) and Ford, Kitsos and Titterington (1989) give a review of optimal design research for nonlinear models. The present paper also indicates that the optimal design methods all have the same characteristics and can be applied to any IRT model. This review not only includes the already known methods but also introduces some alternative selection methods which may prove useful in the future.

First a description of a test design will be given. Then the two most frequently applied information measures will be described, and finally the different criteria for the selection of items for different situations will be reviewed.

Test Design

A test design is characterized by the pattern of the examinee-item combinations. Actually, a test design is connected with a particular test form. For example, a fixed-form test where examinees all take the same items in the test, may be designed in such a way that the items are ordered from very easy to extremely hard. If examinees take the items in the test starting with the most easy item, and stop whenever they give a wrong answer to an item, then the most able examinees will have to answer more items then the examinees with a low ability

level. When examinees are also ordered according to their ability level, then the scores of a test design will approximately have a Guttman scale pattern. In Figure 1 an example of such an approximate Guttman test design is given. The crosses in Figure 1 indicate the examinee-item combinations. The 16 examinees take a test consisting of 20 items. The examinee with the lowest ability level only takes two out of 20 items and the most able examinee takes 19 out of 20 items. It should be emphasized, that the empty cells in the score matrix of the approximate Guttman Test design are empty by design, i.e. the design will determine whether a response is available or not.

Insert Figure 1 about here

Adaptive tests also have special designs. Most adaptive tests are administered in such a way, that each examinee takes a different set of items. The full Nxn matrix of responses of N examinees on a total of n different items will therefore contain a lot of empty cells. In an adaptive test design, the pattern of the cells in the Nxn response matrix is determined by the adaptation process. The design pattern connected with an adaptive test form is certainly not fixed, and may be completely different for examinees having the same ability levels. An example of an adaptive test design is also given in Figure 1. Note that for this particular design an equal number of 10 out of 20 items is administered to each of the 16 examinees.

The designs connected with testlets are more fixed than adaptive test designs. A test containing several testlets will usually have a limited number of paths for an examinee to run though. Depending on their responses to previous

items, examinees may take different items in the testlet and may even take only some of the testlets in the test. The response pattern in a testlet design often follows a kind of branching scheme. Actually, two different types of branching in a testlet design may be distinguished. Examinees may not have to take every testlet in the test. Such a branching may be referred to as *between testlet branching*. When a testlet is structured in such a way that a fixed number of branches of the items within that testlet is made possible, this will be referred to as *within testlet branching*. The third diagram in Figure 1 displays a typical within testlet design connected with a hierarchical testlet (Wainer & Kiely, 1987). In this example, the 16 examinees all take the first item. Then, depending on their response to the first item, they take the second or the third item, and so on. Testlet designs are not as flexible as adaptive test designs, but more flexible than the designs for fixed-form tests.

For the description of a test design some notation is needed. Suppose that we wish to construct an optimal test design for a sample of N examinees (j=1,...,N) and n distinct items (i=1,...,n). Let the matrix $U = \{u_{ij}\}$ represent the response pattern. If the θ-scale with all possible abilities is divided into c distinct categories θ_j, such that $1 \leq c \leq N$, then these categories can be gathered in $\theta' = (\theta_1, \theta_2, \theta_3, \ldots, \theta_c)$, where $\theta \in \mathbb{R}^c$, and $\mathbb{R}^c$ is a c-dimensional set of real numbers. Corresponding with the vector θ is a vector of weights, $W' = (w_1, w_2, w_3, \ldots, w_c)$. These weights can be used in different ways.

The weights in W may be used to characterize the distribution of the sample for which the optimal test is constructed. If, for example, all weights in W are equal, then the sample will have a uniform distribution for the abilities. By a suitable selection of weights a normal ability distribution can also be approximated. The weights can also be used to select only a few θ-levels. If we

wish to find an optimal test design for only two extreme θ-levels, then all but the corresponding two weights w_j will be equal to zero. The weights can also be used to give more weight to certain θ_j-values than to others. Weights can also be used to emphasize the sizes of the intervals between the different θ_j-levels. Some of the criteria discussed in this chapter will make use of such weights.

The items in the test design can be characterized by the vector of structural parameters $\xi' = (\xi_1, \xi_2, \xi_3, \ldots, \xi_n)$, where each element ξ_i may be a vector representing more than one item parameter. For example, for the Rasch model ξ_i will represent the difficulty or location parameter. For extensions of the Rasch model, ξ_i may contain more than one parameter. Of course, items with the same item parameter values may be represented by the same vector ξ_i.

The probability of obtaining a response can now be given by the function $P(\theta_j;\xi_i)$. The mean and variance of the parametric family are $P(\theta_j;\xi_i)$ and $\{P(\theta_j;\xi_i)\ [1- P(\theta_j;\xi_i)]\}$, respectively, and the likelihood function for the data matrix U and θ is:

$$L\ [u;\theta;\xi] = \prod_{j=1}^{c} \prod_{i=1}^{n} P_i(\theta_j;\xi_i)^{w_j\, p_{ij}} \left[1-P_i(\theta_j;\xi_i)\right]^{w_j\,(1-p_{ij})}, \quad (1)$$

where p_{ij} is the proportion of correct responses on item i in category j of θ, and estimation of the parameters $\{\theta_j,\xi_i\}$ can take place by means of the usual maximum likelihood (ML) estimation procedures.

After a model $P(\theta_j;\xi_i)$ is chosen, the test design can be selected. The selection of a test design must be done in such a way, that it will lead to the most accurate estimation of the parameters. The problem, however, is to find

such test designs. More specifically, the problem is to find the set of parameters ξ connected with a certain test form that will enable the most efficient estimation of the parameters in the sample characterized by
$\{\theta, W\}$.

The problem of finding optimal test designs cannot be answered in any general sense and will depend on a number of factors. First, the assumed response model will determine the final outcome. An optimal test design for the Rasch model will generally not be optimal for the two-parameter logistic model. Fortunately, however, the methods for finding optimal test designs can be applied to practically any parametric IRT model.

A second problem is connected with the test form. An optimal test design will differ per test form. For example, an optimal design for a fixed-form test may not be optimal at all for an adaptive test, and vice versa.

Another problem is connected with the parameters themselves. The accuracy of the parameter estimates will depend on the amount of information in the data, and test designs may differ in their amount of information. The variance of the estimators is usually inversely related to the amount of information in the data, and some suitable information measure must be chosen before one can find an optimal test design.

Finally, a selection criterion for the items must be chosen. Since the optimality of a test design will depend on the optimality criterion that was used, the choice of criterion may be crucial. In fact, two alternatives can be distinguished. The first kind of criterion is based on all parameters in $\{\theta, W\}$. This enables a simultaneous optimization procedure for all the parameters in θ. The second kind is formulated on a subset of parameters or even for single parameters, and allows for a stepwise optimization, i.e. for each of the θ_j-values

separately. Although the latter group of criteria has been frequently applied in adaptive testing, these criteria can also be used for the designing of fixed-form tests.

In the following sections the two most frequently applied information measures will be described.

Information Measures

Many different types of information measures for the estimation of parameters have been proposed. Two of the most frequently applied information measures are Fisher's information measure and the Bayesian measure, which is based on the inverse variance of the posterior distribution.

Let the information measure be symbolized by $J(\theta_j)$. Then Fisher's information function connected with the parameter θ_j is defined as:

$$J(\theta_j) \equiv E \{ \frac{\partial}{\partial \theta_j} Log\ L[u;\theta;\xi] \}^2 , \qquad (2)$$

where $L[u;\theta;\xi]$ is the likelihood function. Higher values for $J(\theta_j)$ indicate that more information on the parameter θ_j is available in the sample. Fisher's information has been the most often used measure in test construction. Not only the mathematical programming methods for the construction of fixed-form tests make use of this measure, this measure is also very popular for the construction of adaptive tests.

The second measure is the Bayesian measure. The Bayesian approach to test construction was first proposed by Owen (1975). Instead of using Fisher's information on the ability parameters, Owen (1975) proposed to use the posterior variance. To our knowledge, no mathematical programming procedure based on the maximization of the inverse posterior variance criterion has yet been proposed. To do this, one must first formulate a suitable prior distribution on the abilities being measured. Then, after the selection of response data, the posterior distribution has to be developed by combining the prior distribution with the response data. This means that the use of a Bayesian selection criterion to select items for inclusion in a fixed form test would not be very practical. On the other hand, the implementation of such a Bayesian procedure in the mathematical programming models for two-stage or multi-stage testing procedures proposed by Adema (1990) would be feasible, and it would probably increase the efficiency of the selection procedure, at least when a suitable prior is selected.

When the expected posterior variance is used for item selection, then:

$$J(\theta_j) \equiv E \{ Var^{-1}(\theta_j \mid M(\theta_j)) \} , \qquad (3)$$

where $M(\theta_j)$ is the prior information on θ_j.

For all the parameters in $\{\theta, W\}$, the information measures $J(\theta_j)$ can be grouped into the following vector:

$$J(\theta \mid \xi)' = [\, J(\theta_1),\ J(\theta_2),\ J(\theta_3),\ \ldots\ ,\ J(\theta_c)\,] \,. \qquad (4)$$

This vector $J(\theta \mid \xi)$ contains all available information on the parameters θ in the data and optimality of a test design is usually represented by a function of the two vectors $J(\theta \mid \xi)$ and W. It should be noted, that for multidimensional IRT models the vector θ will become a matrix and $J(\theta \mid \xi)$ will also be a matrix, but the optimality procedures will generally remain the same.

A Class of Optimal Design Criteria

The above given information measures are related to the amount of uncertainty of the estimators of the elements in θ. Optimality of a test design can be defined simultaneously for all the parameters in $\{\theta, W\}$ by considering a function $\Phi(.)$ of $J(\theta \mid \xi)$ and W. Such a simultaneous optimization has the advantage that it will lead to an optimal design for the whole sample of examinees characterized by $\{\theta, W\}$, and also takes into account the shape of its ability distribution.

An optimal test design is a design for which the function Φ $\{J(\theta \mid \xi), W\}$ has the largest possible value, and the problem of finding an optimal test design is actually the problem of maximizing a real-valued concave objective function, i.e.:

$$\text{maximize } \Phi \{J(\theta \mid \xi), W\} \tag{5}$$

subject to

$$\sum w_j = N_{\max} , \tag{6}$$

where N_{max} is some prior specified maximum sample size. In most cases this maximization problem is not easy to solve, and the solution will generally depend on the function $\Phi(.)$ and the information measure being used.

Kiefer (1974) considered a general class of optimality criteria Φ (.) and discussed their properties within an approximate equivalence theory. Members of this class are the so-called product criterion, the sum criterion, and the minimum value criterion. Conceptually, the product criterion can be regarded to correspond with the well-known geometric mean and the sum criterion may be regarded to correspond to the arithmetic mean. This class not only includes these simultaneous optimality criteria, but also includes criteria which are suitable for stepwise optimization. In Table 1 different optimality criteria are displayed, and each of these criteria will be discussed in the following sections.

Insert Table 1 about here

Optimality Criteria for Simultaneous Optimization

Product criterion

The first criterion is a product criterion. The most frequently applied form is the determinant criterion. Usually this criterion is defined as the determinant of an inverse variance-covariance matrix of the estimators, and is often referred to as the D-optimality criterion. This measure was first proposed by Wald (1943), and it is also known as the generalized variance criterion (Anderson, 1984). It can be shown that this criterion is related to Shannon's (1948) information measure of uncertainty about the parameters (Berger, 1991). If the vector $J(\theta \mid \xi)$ represents the main diagonal of a diagonal matrix, then the determinant of that matrix is the product of the main diagonal elements. For an optimal test design the product criterion will become:

$$\Phi \quad \{J(\theta \mid \xi), W\} = \prod_{j=1}^{c} J(\theta_j)^{w_j} . \qquad (7)$$

This criterion has many advantages. Perhaps one of the main reasons for using this criterion is that it has a natural interpretation. It can be shown that it is related to the volume of a confidence region in the parameter space. This means that it can be used to formulate a confidence interval round the parameter estimates. A second feature is that it does not depend on the scale of the independent variable. For the well-known one-, two-, and three-parameter IRT

models, this means that the D-optimality criterion is invariant under linear transformation of the logit scale. Finally, it must be mentioned that its upper bounds for the two-parameter logistic model have been derived by Khan & Yazdi (1988). This means that the actual optimality function value can be compared to the maximal achievable value of the criterion. Such a comparison, for example, was done by Berger (1992b) for two-stage sampling designs.

The D-optimality criterion has also been appealing because of its equivalence with other criteria. The general Equivalence Theorem of Kiefer and Wolfowitz (1960) shows that the D-optimality criterion is equivalent to the G-optimality criterion, which minimizes the maximum variance of the predicted response over the design space. This result indicates that a design is D-optimal if and only if it is G-optimal.

The D-optimality criterion also has some disadvantages. The first disadvantage is that it is generally not sensitive to misspecifications of the model. For example, Abdelbasit & Plankett (1983) showed that for the two-parameter logistic model a D-optimal sampling design for the estimation of the two parameters of a single item consists of only two distinct ability levels. Berger (1992a,b) presents figures of these sampling designs. Because such D-optimal designs are only based on two distinct design points or ability levels, they may not be sensitive to changes in the model specification. Not only minor, but also large deviations in the item characteristic curve may not be detected with data collected according to these designs. Although these problems have been encountered for sampling designs, it can be inferred that these problem will also occur when the D-optimality criterion is applied to test designs.

Another disadvantage of this criterion is that models with a different number of parameters cannot be compared with each other, because the function

depends on the number of parameters being used. It should be noted, however, that this problem also occurs with the other functions.

Sum criterion

A second criterion is the trace or A-optimality criterion. For the test design this function is defined as a weighted sum of information measures connected with the c θ_j- parameters in the sample:

$$\Phi \quad \{J(\theta \mid \xi), W\} = \sum_{j=1}^{c} w_j \, J(\theta_j) \; . \tag{8}$$

This criterion has also often been applied in optimal design research. Although there are cases in which A-optimality is more easily demonstrated than D-optimality, the A-optimality criterion does not have the same advantages as the D-optimality criterion. It is not invariant under linear transformation of the parameter scale and its upper bounds depend on the actual values of the parameters themselves. Although this criterion may seem more appealing to practitioners than the D-optimality criterion, it has hardly been applied in IRT modelling. An example of such a sum criterion for mathematical programming methods has been given by van der Linden and Boekkooi-Timminga (1989).

Minimum value criterion

This criterion may have different forms. Either the minimum value of the information on the parameters is maximized, or the maximum value of the inverse information or asymptotic variance is minimized. An alternative

formulation is based on the smallest eigenvalue of the information matrix, and is referred to as the E-optimality criterion. For IRT models and test designs the smallest value of the vector $J(\theta \mid \xi)$ is maximized:

$$\Phi \quad \{J(\theta \mid \xi), W\} = \min_{j=1}^{c} \{J(\theta_j)\} \ . \tag{9}$$

This criterion is often called a MAXIMIN criterion. An example of a MAXIMIN criterion used as objective function in mathematical programming is given by van der Linden & Boekkooi-Timminga (1989).

Optimality criteria for Stepwise Optimization

The function Φ (.) is defined for the whole set of parameters $\{\theta, W\}$. In some cases, however, optimality for some subset of parameters or for each single parameter may be of interest. For example, a test constructor may want to find a test design that is optimal for the estimation of only the lower ability levels in a sample. Such a selection of the parameters in θ can be established by setting the weights corresponding to the higher θ_j-values equal to zero. The problem is then to find an optimal test design for the subset $\{\theta_s, W_s\}$, where $1 \leq s \leq c$ is the number of parameters in the subset. In many cases, the estimators of the parameters in the subset will not be independent of the estimators of the remaining parameters. In these cases, this dependency should be taken into account when items are selected. The solution to the maximization problem for a

reduced set of parameters is often referred to in the optimal design literature as Φ_S- optimality.

In this section, criteria which are formulated for a single parameter θ_j, i.e. for a single examinee, will be given. These criteria are special cases of the above given criteria for a whole sample of examinees. Instead of a simultaneous maximization, these methods allow for a stepwise optimalization for each single parameter separately. These criteria have been mainly used for the construction of adaptive tests.

In adaptive testing, the construction of a test is individualized for each examinee, and the item selection criterion is formulated for each examinee separately, that is for a single parameter. A distinction between construction methods for fixed-form tests and adaptive tests, is that item selection in adaptive testing is based on previous responses. If an examinee x has an ability θ_x ($x \in N$), then the selection criterion is based on an estimate of the parameter θ_x instead of on the parameter itself. Given such a provisional point estimate, items are selected with the largest information on the ability estimate, i.e.:

$$\Phi \{J(\theta \mid \xi), W\} = J(\hat{\theta}_x). \qquad (10)$$

This criterion was first suggested by Lord (1977) and has been referred to as the maximum information selection criterion (Thissen & Mislevy, 1990). An adaptive test is composed sequentially, after successive administration of the selected items. Compared to the fixed-form test, the adaptive test form may lead to more efficient estimates of the ability, but the stepwise search, for each examinee

separately, through a relatively large number of items will, of course, be more time consuming than for the construction of a fixed-form test.

There is, however, a disadvantage. Criterion (10) is based on a provisional point estimate of θ_x. Especially when the information measure is based on relatively few items, the uncertainty of the estimator may be very high. In these cases the selection of items may be improved by applying a criterion which will take into account the uncertainty of the estimators. Some objective functions that do take into account the uncertainty of the estimators have been proposed by Veerkamp & Berger (1994).

A 100(1-α)% confidence interval for θ_x with lower limit θ_L and upper limit θ_R can be formulated by means of the well-known property that its estimator is asymptotically normally distributed with mean θ_x and variance $J(\theta_x \mid \xi)^{-1}$ which may be replaced by (10). If the pair of vectors $\{\theta_s, W_s\}$ contain all discrete values of the abilities lying within the confidence interval for θ_x, so that the first (lowest) θ_j-value is θ_L and the highest (last) θ_j-value is θ_R, then the area under the information function with limits θ_L and θ_R may be roughly approximated by:

$$\Phi \{J(\theta \mid \xi), W\} = \sum_{j=L}^{R} \omega_j J(\theta_j), \qquad (11)$$

where $\omega_j = |\theta_{j-1} - \theta_j|$. These weights are used to include the size of the intervals between the distinct θ_j-levels, and as such enables approximation of the area under the information function. Item selection in adaptive testing may be improved by applying this interval criterion instead of the maximum (point)

information criterion.

An extension of the interval selection criterion is also possible by including additional weights. If, for example, more weight is given to the information measure $J(\theta \mid \xi)$ when the likelihood is high and less weight is given when the likelihood is low, then a likelihood weighted selection criterion may be formulated as:

$$\Phi \{J(\theta \mid \xi), W\} = \sum_{j=2}^{c} L[u^{(n)};\theta_j;\xi^{(n)}] \quad \omega_j J(\theta_j) , \qquad (12)$$

where $L[u^{(n)};\theta_j;\xi^{(n)}]$ is the likelihood for the responses of the n already administered items. It should be noted, that equations (11) and (12) are equivalent to the weighted sum criterion given in equation (8). Only the weights differ. Some advantages of these criteria are given by Veerkamp & Berger (1994). Because of the additional use of the amount of uncertainty of the estimators, these criteria are expected to perform at least as good as the maximum information criterion. Simulated results given by Veerkamp and Berger (1994) seem to support this conjecture.

An Illustration

One of the main features of simultaneous optimization criteria is that the shape of the ability distribution can be taken into account. An illustration of this feature is presented in Figures 2 and 3. Suppose that we have an item bank with

an infinite number of items. These items have been calibrated by means of the two-parameter IRT model, and cover the full range of combinations of $b_i \in$ <-3,+3> and $a_i \in$ <0.5,3.0>. The product criterion in (7) was used to select the items from the item bank.

In Figure 2 the probability mass functions of the resulting optimal test designs for a positively skewed ability distribution is given for items having three different values of the discrimination parameter a_i = 1.0, 2.0, and 3.0, respectively. These functions indicate that if the items have a discrimination parameter a_i = 1.0, the optimal test would consist of about 80% of the items having difficulty parameter value b_i = -1 and about 20 % of the items with difficulty b_i = -0.5. A very small proportion of items would have a difficulty parameter value b_i = 2.0. When the items have a higher value for a_i, the shape of the probability mass function on the difficulty scale will resemble the positively skewed ability distribution for which the test was designed.

Insert Figure 2 about here

In Figure 3 the optimal test designs are given for a uniform ability distribution. The results in Figure 3 show that for a uniform ability distribution the probability mass functions will also approximately have a uniform shape on the difficulty scale. It should be noted, that the selection of items from the item bank is rather artificially structured, i.e. the items are assumed to have a constant discrimination parameter in each test and exhaustion of the item bank does not play a role, because of the infinite number of items. In this case, the most optimal combination of parameter values can be selected as often as required. For

small item banks, the results will be expected to be quite different.

Insert Figure 3 about here

Discussion and Conclusion

An important aspect of objective measurement in education is the construction of test forms which are not only valid and reliable, but also will produce efficient estimates for the latent trait distribution for which the particular test is designed. In this chapter the designing of different test forms is placed within the general theory of optimal designs. Different methods for optimal design are reviewed in this paper and their properties are discussed. The main conclusion of this paper is that the different test forms, such as fixed-form tests, adaptive tests and testlets, can be constructed by means of comparable methods, and that these methods are actually the same as the procedures which have been used in optimal design theory.

The construction of an optimal test design can be viewed as an optimization problem, and several algorithms for finding optimal designs have been proposed in the literature. Among those optimization procedures are the mathematical programming procedures, which have been used for the construction of fixed-form tests by Adema (1990) and Boekkooi (1989), among others. Apart from these procedures several other optimal design algorithms have been applied in other fields of research. See for example, Cook and Nachtsheim (1980) for a review. Perhaps the most promising algorithms for test construction

are the sequential design algorithms. These procedures have been studied extensively by Ford, Titterington and Wu (1985), Wu (1985), Wu and Wynn (1978) and Wynn (1970), and were applied to IRT modelling by Berger (1992ab, in press). The sequential construction of optimal test designs by means of the methods discussed in this paper is straightforward.

References

Abdelbasit, K.M. & Plankett, R.L. (1983). Experimental design for binary data. Journal of the American Statistical Association, 78, 90-98.

Adema, J.J.(1990) Models and algorithms for the construction of achievement tests. Ph.D. thesis, University of Twente.

Anderson, T.W. (1984). An introduction to multivariate statistical analysis (2nd ed.) New York: Wiley.

Berger, M.P.F. (1991). On the efficiency of IRT models when applied to different sampling designs. Applied Psychological Measurement, 15,293-306.

Berger, M.P.F. (1992a). Generation of optimal designs for nonlinear models when the design points are incidental parameters. In: Y. Dodge and J. Whittaker (Eds.) Computational Statistics, Volume 2 (pp. 200-208), New York: Springer Verlag.

Berger, M.P.F. (1992b). Sequential sampling designs for the two-parameters item response theory model. Psychometrika, 57, 521-538.

Berger, M.P.F. (in press). D-optimal sequential sampling designs for item response theory models. Journal of Educational Statistics.

Berger, M.P.F. & van der Linden, W.J. (1992). Optimality of sampling designs in item response theory models. In: M. Wilson (Ed.), Objective measurement: Theory into practice (pp.274-288). Norwood, NJ: Ablex Publishing Company.

Boekkooi-Timminga, E. (1989). Models for computerized test construction, Ph.D. Thesis, University of Twente.

Boekkooi-Timminga,E. & Sun, L. (1991). Contest: A computerized test construction system. In: Hoogstraten, J. & van der Linden, W.J. (Eds.), Methodologie: Onderwijsresearchdagen 91 (pp. 69-77). Amsterdam: SCO.

Cook, R.D. & Nachtsheim, C.J. (1980). A comparison of algorithms for constructing exact D-optimal designs. Technometrics, 22, 315-324.

Ford, I., Titterington, D.M. and Wu, C.F.J. (1985). Inference and sequential design. Biometrika, 72, 545-551.

Ford, I, Kitsos, C.P. and Titterington, D.M. (1989). Recent advances in nonlinear experimental design. Technometrics, 31,49-60.

Khan, M.K. & Yazdi, A.A. (1988). On D-optimal designs. Journal of Statistical Planning and Inference, 18, 83-91.

Kiefer, J. (1974). General equivalence theory for optimum designs (Approximate theory). The Annals of Statistics, 2, 849-879.

Kiefer, J. & Wolfowitz, J. (1960). The equivalence of two extremum problems. Canadian Journal of Mathematics, 30, 271-319.

Lord, F.M. (1971) Robbins-Monro procedures for tailored testing. Educational and Psychological Measurement, 31, 3-31.

Lord, F.M. (1980). Applications of item response theory to practical testing problems, Hillsdale, NJ: Erlbaum.

Mislevy, R.J., & Wu, P.K. (1988). Inferring examinee ability when some item responses are missing (Research Report 88-48-ONR). Princeton, NJ: Educational Testing service.

Owen, R.J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. Journal of the American Statistical Association, 70, 351-356.

Shannon, C.E. (1948). A mathematical theory of communication. Bell System Technical Journal, 27, 379-423, 623-656.

Silvey, S.D. (1980). Optimal design. London: Chapman & Hall.

Theunissen, T.J.J.M. (1986). Binary programming and test design. Psychometrika, 50, 411-420.

Thissen, D. and Mislevy, R.J. (1990). Testing algorithms. In: Wainer, H. (Ed.) Computerized adaptive testing: A primer (pp. 103-135). Hillsdale, NJ: Lawrence Erlbaum Ass.

Van der Linden, W.J. & Boekkooi-Timminga, E. (1989). A maximin model for test design with practical constraints. Psychometrika, 53, 237-247.

Veerkamp, W.J.J. & Berger, M.P.F. (1994). A comparison of different item selection criteria for adaptive testing. Research Report, Enschede: University of Twente, Faculty of Educational Science and Technology.

Verschoor, A. (1991). OTD: Optimal Test Design, (Computerprogram). Arnhem:CITO.

Wainer, H. (1990). Computerized adaptive testing: A primer. Hillsdale, NJ: Lawrence Erlbaum Ass.

Wainer, H. & Kiely, G.L. (1987). Item clusters and computerized adaptive testing: A case for testlets. Journal of Educational Measurement, 24, 185-202.

Wainer, H. & Lewis, C. (1990). Towards a psychometrics for testlets. Journal of Educational Measurement, 27, 1-14.

Wald, A. (1943). On the efficient design of statistical investigations. Annals of Mathematical Statistics, 14, 134-140.

Weiss, D.J. (1976). Adaptive testing research in Minnesota: Overview, recent results, and future directions. In C.L. Clark (Ed.), Proceedings of the First Conference on Computerized Adaptive Testing (pp.24-35). Washington, DC: United States Civil Service Commission.

Weiss, D.J. (1978). Proceedings of the 1977 Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota.

Wu, C.F.J. (1985). Asymptotic inference from sequential design in nonlinear situation, Biometrika, 72, 533-558.

Wu, C.F.J. and Wynn, H.P. (1978). The convergence of general step-length algorithms for regular optimum design criteria. The Annals of Statistics, 6, 1273-1285.

Wynn, H.P. (1970). The sequential generation of D-optimum experimental designs. Annals of Mathematical Statistics, 41, 1655-1664.

Table 1

Different Item Selection Criteria for
Simultaneous and Stepwise Optimization

Simultaneous Optimization	Stepwise Optimization
Product Criterion	Maximum Information Criterion
Sum Criterion	Interval Information Criterion
Min. Value Criterion	Weighted Interval Criterion

Subject Index

- adaptive testing
- testlets
- fixed-form test
- information measure function
- mathematical programming
- (optimal) test design / test construction
- item bank
- item selection
- optimal sampling
- parameter estimation
- Guttman scale
- (maximum) likelihood
- Rasch model/(two-parameter) logistic model
- IRT
- Bayesian measure
- posterior variance / distribution
- prior distribution
- product criterion
- (weighted) sum criterion
- minimum value criterion
- stepwise optimization
- simultaneous optimization

Subject Index (vervolg)

- determinant criterion
- D-optimality criterion
- generalized variance criterion
- Shannon's information measure
- general Equivalence Theorem
- G-optimality criterion
- trace criterion
- A-optimality criterion
- E-optimality criterion
- maximin criterion
- Φ_s-optimality
- maximum information selection criterion
- interval criterion selection criterion
- weighted interval selection criterion
- discrimination parameter
- difficulty parameter
- ability distribution

Figure Captions

Figure 1 Three test designs.

Figure 2 Simultaneously designed optimal test design for a positively skewed ability distribution.

Figure 3 Simultaneously designed optimal test design for a uniform ability distribution.

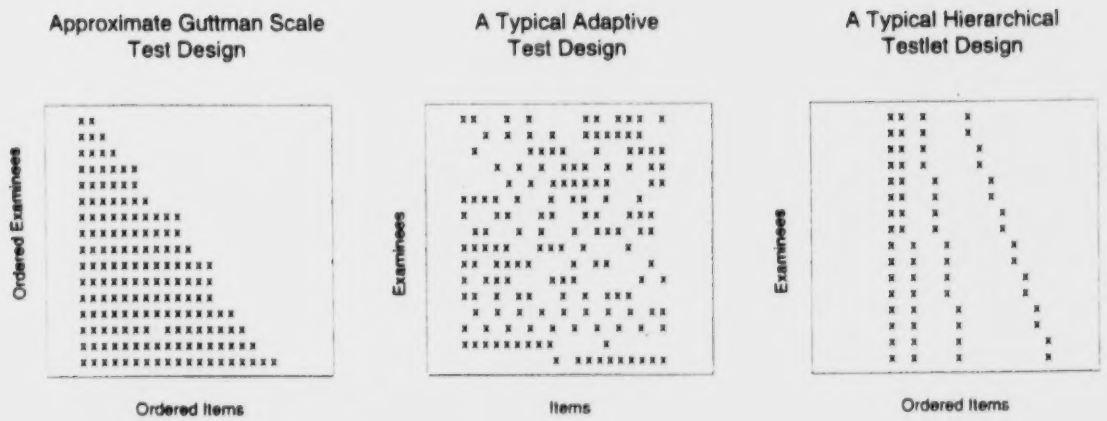

Figure 1

BEST COPY AVAILABLE

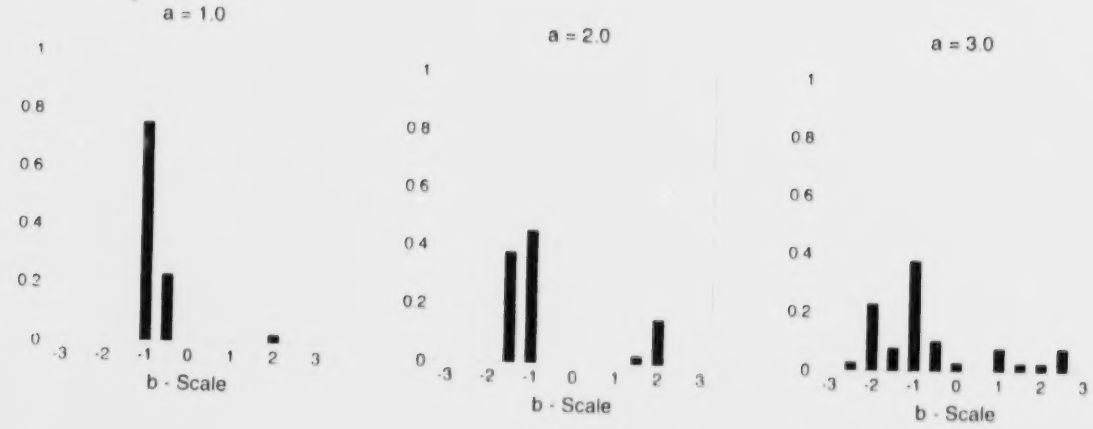

Figure 2

BEST COPY AVAILABLE

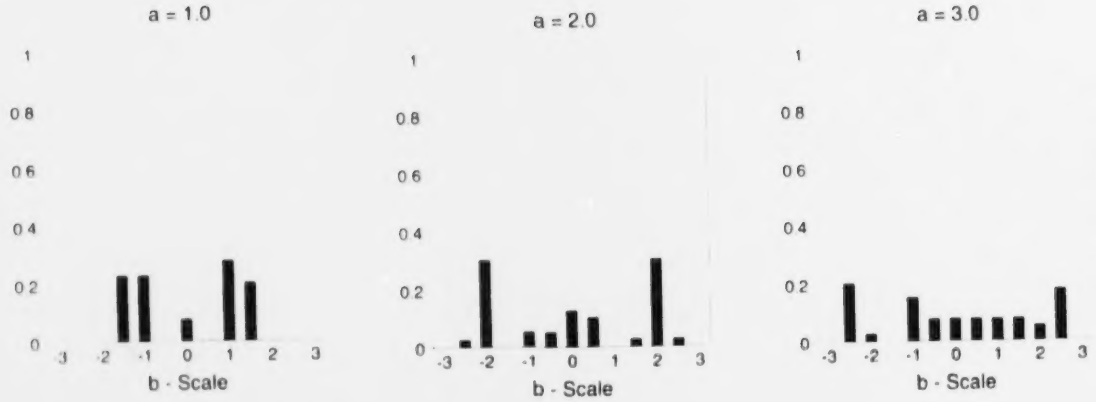

Figure 3

BEST COPY AVAILABLE

Titles of recent Research Reports from the Department of Educational Measurement and Data Analysis. University of Twente, Enschede, The Netherlands.

RR-94-4 M.P.F. Berger & W.J.J. Veerkamp, *A review of selection methods for optimal design*

RR-94-3 W.J. van der Linden, *A conceptual analysis of standard setting in large-scale assessments*

RR-94-2 W.J. van der Linden & H.J. Vos, *A compensatory approach to optimal selection with mastery scores*

RR-94-1 R.R. Meijer, *The influence of the presence of deviant item score patterns on the power of a person-fit statistic*

RR-93-1 P. Westers & H. Kelderman, *Generalizations of the Solution-Error Response-Error Model*

RR-91-1 H. Kelderman, *Computing Maximum Likelihood Estimates of Loglinear Models from Marginal Sums with Special Attention to Loglinear Item Response Theory*

RR-90-8 M.P.F. Berger & D.L. Knol, *On the Assessment of Dimensionality in Multidimensional Item Response Theory Models*

RR-90-7 E. Boekkooi-Timminga, *A Method for Designing IRT-based Item Banks*

RR-90-6 J.J. Adema, *The Construction of Weakly Parallel Tests by Mathematical Programming*

RR-90-5 J.J. Adema, *A Revised Simplex Method for Test Construction Problems*

RR-90-4 J.J. Adema, *Methods and Models for the Construction of Weakly Parallel Tests*

RR-90-2 H. Tobi, *Item Response Theory at subject- and group-level*

RR-90-1 P. Westers & H. Kelderman, *Differential item functioning in multiple choice items*

Research Reports can be obtained at costs from Bibliotheek, Faculty of Educational Science and Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.

faculty of
EDUCATIONAL SCIENCE
AND TECHNOLOGY

A publication by
The Faculty of Educational Science and Technology of the University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands